# <論文・報告>Efficient Filtering Algorithm with Fused Constraint

AUTHOR(S):

Ono, Tasuku

CITATION:

Ono, Tasuku. <論文・報告>Efficient Filtering Algorithm with Fused Constraint. ELCAS Journal 2020, 5: 26-28

ISSUE DATE:

2020-04

URL:

http://hdl.handle.net/2433/251398

# Efficient Filtering Algorithm with Fused Constraint

Tasuku Ono

Koyo Gakuin High School

## Abstract

Sparse modeling is an effective machine-learning method to analyze missing and sparse data. The fused LASSO (least absolute shrinkage and selection operator) is one of the techniques used in sparse modeling. In this paper, we propose a simple yet effective optimization technique for the fused LASSO problem based on the method [1]. We then show the main advantage of the proposed technique: the proposed update formula monotonically decreases the objective function.

## 1 Introduction

Improving technology in experimental instruments and data processing today makes it possible to deal with a large amount of data in the field of natural science. Machine learning is one of the effective methods to analyze and examine mass data. For example, machine learning methods have contributed to first-ever reconstruction of the image of a black hole shadow from the massive data sets collected by the Event Horizon Telescope [2].

One of the key challenges to handle data in natural science is that observed data tends to be missing and sparse. A machine-learning technique called *sparse modeling* is effective to handle such data due to its capability of identifying important variables automatically while reconstructing the data. The least absolute shrinkage and selection operator (LASSO) [3], which is the least-square regression with $\ell_1$ regularization, is the most widely used sparse modeling technique. LASSO is a feature selection method and it assumes all the input variables are independent. The LASSO optimization problem for regression problem for matrix is written as follows:

$$\min_{\boldsymbol{w}} \sum_{i=1}^{n} (y_i - \boldsymbol{w}^\top \boldsymbol{x}_i)^2 + \lambda \sum_{i=1}^{d} |\, w_i \,|,$$

where $\boldsymbol{w} = [w_1, w_2, ..., w_d]^\top \in \mathbb{R}^d$ and $\lambda \geq 0$ is the regularization parameter. LASSO can set some regression parameters of the explanatory variables to 0, while it solves the optimization problem.

Fused LASSO is an additional version of LASSO, where the sum of the absolute value of the differences between adjacent coefficients is added as a sparse regularization term [4]. Fused LASSO optimization problem for matrix can be given as follows:

$$\min_{\boldsymbol{w}} \sum_{i=1}^{n} (y_i - \boldsymbol{w}^\top \boldsymbol{x}_i)^2 + \lambda_1 \sum_{i=1}^{d} |\, w_i \,| + \lambda_2 \sum_{i=2}^{d} |\, w_i - w_{i-1} \,|,$$

where $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$ are regularization parameters. Fused LASSO allows to specify the regression parameters which have the same degree of contribution to the explanatory variables. The fused LASSO is in particular useful for image denoising and the change point detection. Recently, it has been used for constructing a black hole image using the total variation and the $\ell_1$ regularizer [2].

In this paper, we propose a simple yet effective optimization technique for the total variation and the $\ell_1$ regularization. More specifically, we employ the majorization-minimization based technique [1] and derive the closed-form update formula for the fused LASSO problem. The key advantage of the proposed optimization technique is that it can show the update monotonically decrease the objective function.

## 2 Problem Formulation

The notations used in this paper are described as follows. Scalars, vectors, matrices are written as lightface lowercase letters, boldface lowercase letters, boldface uppercase letters, respectively. The $n$-th element of a vector $\boldsymbol{x}$ is denoted by $x_n$. As for a matrix $\boldsymbol{X}$, the $(i,j)$ element of the matrix is denoted by $x_{i,j}$.

- $x$: scalar
- $\boldsymbol{x} = (x_1, x_2, ..., x_m)^\top \in \mathbb{R}^m$: vector
- $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n) = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m,1} & x_{m,2} & \cdots & x_{m,n} \end{bmatrix} \in \mathbb{R}^{m \times n}$: matrix

In this paper, we consider the matrix denoising problem. More specifically, with a given noisy matrix $\boldsymbol{X} \in \mathbb{R}^{m \times n}$, we recover a clean matrix $\boldsymbol{W}$.

onotask@gmail.com (Tasuku Ono, Koyo Gakuin High School)
myamada@i.kyoto-u.ac.jp (Makoto Yamada, Kyoto University)

## 3 Proposed Method

To obtain clean matrix $\boldsymbol{W}$, we solve the following optimization problem:

$$\min_{\boldsymbol{W}} \; J(\boldsymbol{W}) \tag{1}$$

where

$$J(\boldsymbol{W}) = \sum_{i=1}^{m}\sum_{j=1}^{n}(x_{i,j}-w_{i,j})^2 + \lambda_1\sum_{i=1}^{m}|w_i| + \lambda_2\sum_{i=1}^{m-1}\sum_{j=1}^{n-1}\sqrt{(w_{i+1,j}-w_{i,j})^2+(w_{i,j+1}-w_{i,j})^2} \tag{2}$$

and $w_{i,j}$ and $x_{i,j}$ are $(i,j)$ elements of the matrices, $\boldsymbol{W}$ and $\boldsymbol{X}$, respectively:

$$w_{i,j} = [\boldsymbol{W}]_{i,j} \text{ and } x_{i,j} = [\boldsymbol{X}]_{i,j}$$

For the purpose of simplifying the optimization problem, we remove the lasso regularization term $\lambda_1\sum_{i=1}^{m}|w_i|$ from the original cost function and rewrite it as follows:

$$J(\boldsymbol{W}) = \sum_{i=1}^{m}\sum_{j=1}^{n}(x_{i,j}-w_{i,j})^2 + \lambda\sum_{i=1}^{m-1}\sum_{j=1}^{n-1}\sqrt{(w_{i+1,j}-w_{i,j})^2+(w_{i,j+1}-w_{i,j})^2} \tag{3}$$

Thus, the cost function $J(\boldsymbol{W})$ is written in terms of matrices as follows:

$$J(\boldsymbol{W}) = \|\boldsymbol{X}-\boldsymbol{W}\|_{\text{Frob}}^2 + \lambda\|\boldsymbol{W}\|_{\text{TV}}, \tag{4}$$

where $\|\boldsymbol{W}\|_{\text{Frob}}$, the Frobenius norm of the matrix $\boldsymbol{W}$, is defined as follows:

$$\|\boldsymbol{W}\|_{\text{Frob}} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}w_{i,j}^2}.$$

However, TV term is not differentiable at 0 so that the optimization will be difficult. Therefore, instead of minimizing the original loss function, we minimize the alternative loss function:

$$\sum_{i=1}^{m-1}\sum_{j=1}^{n-1}\frac{(w_{i+1,j}-w_{i,j})^2+(w_{i,j+1}-w_{i,j})^2}{\sqrt{(w_{i+1,j}^{(t)}-w_{i,j}^{(t)})^2+(w_{i,j+1}^{(t)}-w_{i,j}^{(t)})^2+\epsilon}}, \tag{5}$$

where $\epsilon > 0$.

We reformulate this expression by using matrices in order to get the update formula easily. The reformulated TV term can be written as

$$\sum_{i=1}^{m-1}\sum_{j=1}^{n-1}\frac{(w_{i+1,j}-w_{i,j})^2+(w_{i,j+1}-w_{i,j})^2}{\sqrt{(w_{i+1,j}^{(t)}-w_{i,j}^{(t)})^2+(w_{i,j+1}^{(t)}-w_{i,j}^{(t)})^2+\epsilon}} = \sum_{i=1}^{m-1}\sum_{j=1}^{n-1}\frac{\|\boldsymbol{D}_{i,j}\text{vec}(\boldsymbol{W})\|_2^2}{\|\boldsymbol{D}_{i,j}\text{vec}(\boldsymbol{W}^{(t)})\|_{2,\epsilon}}, \tag{6}$$

where $\boldsymbol{D}_{i,j} \in \mathbb{R}^{2\times nm}$ is a matrix which takes the difference between its adjacent elements, $\boldsymbol{W}^{(t)}$ is a $t$-th updated matrix of the clean matrix $\boldsymbol{W}$, vec$(\cdot)$ is the vectorization operator (i.e. vec$(\boldsymbol{W}) \in \mathbb{R}^{mn}$, and $\|\boldsymbol{a}\|_{2,\epsilon} = \|\boldsymbol{a}\|_{2,\epsilon} = \sqrt{\boldsymbol{a}^\top\boldsymbol{a}+\epsilon}$.

As an example, consider a 2 by 2 matrix case. We can write down the differences between $w_{2,1}$ and $w_{1,1}$, and that between $w_{1,2}$ and $w_{1,1}$ as follows:

$$w_{2,1}-w_{1,1} = \begin{bmatrix}-1 & 0 & 1 & 0\end{bmatrix}\begin{bmatrix}w_{1,1}\\w_{1,2}\\w_{2,1}\\w_{2,2}\end{bmatrix}$$

and

$$w_{1,2}-w_{1,1} = \begin{bmatrix}-1 & 1 & 0 & 0\end{bmatrix}\begin{bmatrix}w_{1,1}\\w_{1,2}\\w_{2,1}\\w_{2,2}\end{bmatrix},$$

where $w_{1,1}$ $(i,j) \in 1, 2$ is an element of the matrix $\boldsymbol{W}$. Thus we have

$$(w_{2,1}-w_{1,1})^2 + (w_{1,2}-w_{1,1})^2 = \|\boldsymbol{D}_{1,1}\text{vec}(\boldsymbol{W})\|_2^2 \tag{7}$$

where

$$\boldsymbol{D}_{1,1} = \begin{bmatrix}-1 & 0 & 1 & 0\\-1 & 1 & 0 & 0\end{bmatrix}, \text{vec}(\boldsymbol{W}) = \begin{bmatrix}w_{1,1}\\w_{1,2}\\w_{2,1}\\w_{2,2}\end{bmatrix}$$

From the above, the alternative loss function is written as

$$\tilde{J}(\boldsymbol{W}) = \|\text{vec}(\boldsymbol{X})-\text{vec}(\boldsymbol{W})\|_2^2 + \lambda\sum_{i=1}^{m-1}\sum_{j=1}^{n-1}\frac{\|\boldsymbol{D}_{i,j}\text{vec}(\boldsymbol{W})\|_2^2}{\|\boldsymbol{D}_{i,j}\text{vec}(\boldsymbol{W}^{(t)})\|_{2,\epsilon}}. \tag{8}$$

Therefore, the optimization problem can be rewritten as follows:

$$\min_{\boldsymbol{W}} \; \tilde{J}(\boldsymbol{W})$$

In order to solve this optimization problem, we differentiate $\tilde{J}(\boldsymbol{W})$ with respect to $\boldsymbol{W}$:

$$\begin{aligned}\frac{\partial\tilde{J}(\boldsymbol{W})}{\partial\boldsymbol{W}} &= \frac{\partial}{\partial\boldsymbol{W}}\|\text{vec}(\boldsymbol{X})-\text{vec}(\boldsymbol{W})\|_2^2 \\ &\quad + \left(\lambda\sum_{i=1}^{m-1}\sum_{j=1}^{n-1}\frac{1}{\|\boldsymbol{D}_{i,j}\text{vec}(\boldsymbol{W}^{(t)})\|_{2,\epsilon}}\right)\frac{\partial}{\partial\boldsymbol{W}}\|\boldsymbol{D}_{i,j}\text{vec}(\boldsymbol{W})\|_2^2 \\ &= -2(\text{vec}(\boldsymbol{X})-\text{vec}(\boldsymbol{W})) \\ &\quad + 2\lambda\sum_{i=1}^{m-1}\sum_{j=1}^{n-1}\frac{1}{\|\boldsymbol{D}_{i,j}\text{vec}(\boldsymbol{W}^{(t)})\|_{2,\epsilon}}\boldsymbol{D}_{i,j}^\top\boldsymbol{D}_{i,j}\text{vec}(\boldsymbol{W})\end{aligned}$$

The updated matrix $\boldsymbol{W}^{(t+1)}$ is given by solving the following equation:

$$\frac{\partial\tilde{J}(\boldsymbol{W})}{\partial\boldsymbol{W}} = 0 \tag{9}$$

We can substitute $\frac{\partial\tilde{J}(\boldsymbol{W})}{\partial\boldsymbol{W}}$ by the equation above and solve it as follows:

$$-2(\text{vec}(\boldsymbol{X})-\text{vec}(\boldsymbol{W}^{(t+1)})) + 2\lambda\sum_{i=1}^{m-1}\sum_{j=1}^{n-1}\frac{1}{\|\boldsymbol{D}_{i,j}\text{vec}(\boldsymbol{W}^{(t)})\|_{2,\epsilon}}\boldsymbol{D}_{i,j}^\top\boldsymbol{D}_{i,j}\text{vec}(\boldsymbol{W}^{(t+1)}) = 0$$

$$\left(\boldsymbol{I}_{nm}+\lambda\sum_{i=1}^{m-1}\sum_{j=1}^{n-1}\frac{1}{\|\boldsymbol{D}_{i,j}\mathrm{vec}(\boldsymbol{W}^{(t)})\|_{2,\epsilon}}\boldsymbol{D}_{i,j}^{\top}\boldsymbol{D}_{i,j}\right)\mathrm{vec}(\boldsymbol{W}^{(t+1)})=\mathrm{vec}(\boldsymbol{X})$$

$$\mathrm{vec}(\boldsymbol{W}^{(t+1)})=\left(\boldsymbol{I}_{nm}+\lambda\sum_{i=1}^{m-1}\sum_{j=1}^{n-1}\frac{1}{\|\boldsymbol{D}_{i,j}\mathrm{vec}(\boldsymbol{W}^{(t)})\|_{2,\epsilon}}\boldsymbol{D}_{i,j}^{\top}\boldsymbol{D}_{i,j}\right)^{-1}\mathrm{vec}(\boldsymbol{X}) \tag{10}$$

where $\boldsymbol{I}_{nm}$ is a $n \times m$ identity matrix.

Next, we are going to show that minimizing the alternative cost function $\tilde{J}(\boldsymbol{W})$ is the global optimum problem of minimizing the original cost function $J(\boldsymbol{W})$. In order to prove it, we need the following lemma:

**Lemma 1** *For any nonzero vectors, $\boldsymbol{w}$, $\boldsymbol{w}^{(t)} \in \mathbb{R}^m$, the following holds*

$$\|\boldsymbol{w}\|_{2,\epsilon}-\frac{\|\boldsymbol{w}\|_{2,\epsilon}^2}{2\|\boldsymbol{w}^{(t)}\|_{2,\epsilon}}\leq\|\boldsymbol{w}^{(t)}\|_{2,\epsilon}-\frac{\|\boldsymbol{w}^{(t)}\|_{2,\epsilon}^2}{2\|\boldsymbol{w}^{(t)}\|_{2,\epsilon}} \tag{11}$$

*proof.* Two scalar values $u+\epsilon$ and $u_t+\epsilon$ satisfy the following inequality:

$$\left(\sqrt{u+\epsilon}-\sqrt{u_t+\epsilon}\right)^2\geq 0$$

Therefore, we have

$$\begin{aligned}\left(\sqrt{w+\epsilon}-\sqrt{w_t+\epsilon}\right)^2\geq 0 &\Rightarrow w_t+\epsilon\geq 2\sqrt{w+\epsilon}\sqrt{w_t+\epsilon}-(w+\epsilon)\\ &\Rightarrow\frac{w_t+\epsilon}{2\sqrt{w_t+\epsilon}}\geq\sqrt{w+\epsilon}-\frac{w+\epsilon}{2\sqrt{w_t+\epsilon}}\\ &\Rightarrow\sqrt{w_t+\epsilon}-\frac{w_t+\epsilon}{2\sqrt{w_t+\epsilon}}\\ &\quad\geq\sqrt{w+\epsilon}-\frac{w+\epsilon}{2\sqrt{w_t+\epsilon}}\end{aligned}$$

We arrive at the Eq.(11) by substituting the $w+\epsilon$ and $w_t+\epsilon$ by $\|\boldsymbol{w}\|_{2,\epsilon}$ and $\|\boldsymbol{w}^{(t)}\|_{2,\epsilon}$ respectively. □

In Lemma 1, we prove that any nonzero vectors $\boldsymbol{w}$, $\boldsymbol{w}^{(t)} \in \mathbb{R}^m$ satisfy the Eq.(11). By substituting $\boldsymbol{D}_{i,j}\mathrm{vec}(\boldsymbol{W})$ and $\boldsymbol{D}_{i,j}\mathrm{vec}(\boldsymbol{W}^{(t)})$ for $\boldsymbol{w}$ and $\boldsymbol{w}^{(t)}$ respectively, we have the following inequality:

$$\begin{aligned}\|\boldsymbol{D}_{i,j}\mathrm{vec}(\boldsymbol{W})\|_{2,\epsilon}-\frac{\|\boldsymbol{D}_{i,j}\mathrm{vec}(\boldsymbol{W})\|_{2,\epsilon}^2}{2\|\boldsymbol{D}_{i,j}\mathrm{vec}(\boldsymbol{W}^{(t)})\|_{2,\epsilon}}&\\ \leq\|\boldsymbol{D}_{i,j}\mathrm{vec}(\boldsymbol{W}^{(t)})\|_{2,\epsilon}-\frac{\|\boldsymbol{D}_{i,j}\mathrm{vec}(\boldsymbol{W}^{(t)})\|_{2,\epsilon}^2}{2\|\boldsymbol{D}_{i,j}\mathrm{vec}(\boldsymbol{W}^{(t)})\|_{2,\epsilon}}&\end{aligned} \tag{12}$$

When we solve the alternative optimization problem, the following inequality holds:

$$\frac{\|\boldsymbol{D}_{i,j}\mathrm{vec}(\boldsymbol{W})\|_{2,\epsilon}^2}{\|\boldsymbol{D}_{i,j}\mathrm{vec}(\boldsymbol{W}^{(t)})\|_{2,\epsilon}}\leq\frac{\|\boldsymbol{D}_{i,j}\mathrm{vec}(\boldsymbol{W}^{(t)})\|_{2,\epsilon}^2}{\|\boldsymbol{D}_{i,j}\mathrm{vec}(\boldsymbol{W}^{(t)})\|_{2,\epsilon}} \tag{13}$$

From the Eq.(12) and the Eq.(13), we have:

$$\|\boldsymbol{D}_{i,j}\mathrm{vec}(\boldsymbol{W})\|_{2,\epsilon}\leq\|\boldsymbol{D}_{i,j}\mathrm{vec}(\boldsymbol{W}^{(t)})\|_{2,\epsilon} \tag{14}$$

That is to say, once we minimize the alternative cost function $\tilde{J}(\boldsymbol{W})$, we simultaneously minimize the original cost function $J(\boldsymbol{W})$ in the sense of global optimization. This shows that we can use the proposed loss function in the Eq.(8) as an alternative for the original loss function in the Eq.(2) to solve the proposed optimization problem.

## 4 Conclusion

In this paper, we proposed a simple but effective optimization technique for the Fused LASSO regression by considering the matrix reconstruction problem. The proposed technique and its update formula rest on the majorization-minimization based technique [1]. Moreover, the proposed technique has a major advantage in that the update formula monotonically decreases the original objective function. Applying the proposed technique to real image reconstruction problem will possibly enable us to investigate its effectiveness.

## Acknowledgements

The author would like to express my sincere appreciation to Professor Yamada and members in his laboratory for their assistance.

## References

[1] Feiping Nie, Heng Huang, Xiao Cai, and Chris H Ding. Efficient and robust feature selection via joint l2, 1-norms minimization. In *Advances in neural information processing systems*, pages 1813–1821, 2010.

[2] Kazunori Akiyama, Kazuki Kuramochi, Shiro Ikeda, Vincent L Fish, Fumie Tazaki, Mareki Honma, Sheperd S Doeleman, Avery E Broderick, Jason Dexter, Monika Mościbrodzka, et al. Imaging the schwarzschild-radius-scale structure of m87 with the event horizon telescope using sparse modeling. *The Astrophysical Journal*, 838(1):1, 2017.

[3] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[4] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.